Polysémie lexicale

Guillaume Jacquet, Fabienne Venant, Bernard Victorri

Sémantique et traitement automatique du langage naturel, Hermès, 2005 – Version préliminaire

Polysémie lexicale

Guillaume Jacquet, Fabienne Venant, Bernard Victorri

1. Introduction

La polysémie lexicale représente une difficulté majeure pour tous les modèles de représentation et de calcul du sens. Quels que soient la définition du sens et le type de représentation que l'on adopte pour les unités lexicales, on se trouve en effet confronté d'emblée au problème de circularité engendré par l'omniprésence de cette polysémie. En effet, puisque la plupart des unités lexicales peuvent prendre des sens différents, leur sens dans une phrase donnée dépend du sens de cette phrase, et, bien entendu, le sens de la phrase est lui-même fonction du sens des unités qui la composent. On ne peut donc pas ignorer la polysémie sous peine de tomber dans un cercle vicieux, quel que soit le cadre théorique que l'on adopte. D'un point de vue pratique, cela signifie qu'il faut inclure, dans les systèmes de traitement automatique qui ont besoin de calculer le sens lexical, des mécanismes généraux de désambiguïsation, qui doivent prendre en compte le co-texte immédiat et, dans la mesure du possible, le contexte au sens large (y compris la situation et les conditions d'énonciation), pour calculer de manière plus ou moins dynamique le sens de chaque occurrence des unités polysémiques du texte que l'on doit traiter.

Cette tâche de désambiguïsation peut être plus ou moins lourde, selon le type de texte auquel on a affaire et le type de résultat qui est visé. Mais comme l'avait montré Pitrat ([PIT 85], p. 14-16) il y a déjà longtemps à propos de textes de commentaires de parties d'échec, même des textes dédiés à des sujets très précis sont loin d'être exempts de termes polysémiques utilisés dans des sens différents et assez imprévisibles. Contrairement à l'affirmation de Rastier ([RAS 94], p. 51), la polysémie n'est pas un simple « artefact de la linguistique » qui résulterait de l'absence de prise en compte du niveau textuel dans la description du sens des unités[1].

Cela dit, dans des applications très ciblées, comme l'extraction d'informations précises (cf. chap. 8) sur des corpus spécialisés, avec des thématiques assez bien délimitées, la tâche de désambiguïsation va être à l'évidence plus légère que dans des applications plus « généralistes », comme l'indexation automatique (cf. chap. 9) de documents susceptibles d'évoquer les sujets les plus divers. Prenons l'exemple du verbe *rouler*, que nous avons déjà évoqué (chap. 2, § 1.3) et que nous allons discuter plus en détail ci-dessous au § 2.3. Dans un corpus de constats d'accidents de voiture, ce verbe sera massivement utilisé dans le sens 'se déplacer dans un véhicule à roues'. Même si l'on ne peut pas exclure totalement d'autres sens du verbe dans ce type de contexte, comme 'se déplacer en tournant sur soi-même' (*Le motocycliste a été projeté et il a roulé dans le fossé*), voire 'duper' (*Le camionneur m'a roulé en me faisant croire que j'étais dans mon tort*), il est clair que ces emplois seront suffisamment rares pour ne pas augmenter significativement un taux d'erreur si l'on décidait de le traiter comme un verbe monosémique. Le problème se pose de manière radicalement différente si l'on a affaire à un corpus généraliste, par exemple un quotidien ou le Web : tous les sens (plus d'une vingtaine) sont susceptibles d'être

1. Pour une discussion plus approfondie de ce point, voir [VIC 97].

rencontrés, et une enquête rapide sur le Web francophone montre qu'au moins six sens très différents sont présents à des fréquences massives[2].

Comme nous le verrons, le terme de polysémie recouvre en fait des phénomènes extrêmement variés. Certaines formes sont très systématiques : elles peuvent être décrites par des règles générales assez simples et elles s'appliquent à tous les éléments lexicaux qui remplissent les critères énoncés par ces règles. D'autres sont au contraire très spécifiques de telle ou telle unité lexicale, résultat de la singularité de son histoire individuelle. Entre ces deux extrêmes, on trouve toute une gamme de phénomènes qui dépassent le cadre de l'unité individuelle, mais qui ne se laissent pas non plus décrire par une règle systématique, ou, plus précisément, qu'il est difficile de résumer par une règle synthétique suffisamment simple. Pour un même mécanisme, il faut aussi distinguer les phénomènes discursifs, créations éphémères de sens nouveaux au fil du discours, des phénomènes lexicalisés, qui se sont inscrits de manière durable et conventionnelle dans la langue.

A cette diversité des formes de polysémie correspondent des stratégies de traitements automatiques tout aussi hétérogènes. Nous allons commencer par faire un inventaire de ces phénomènes, de manière à prendre la mesure des problèmes que se posent linguistes et modélisateurs. Dans un deuxième temps, nous présenterons un modèle théorique particulier de la polysémie lexicale, développé par notre équipe, qui met l'accent sur une notion de proximité entre les sens d'une unité polysémique et sur le caractère dynamique du calcul du sens. En effet, dans cette approche, les différents sens d'une unité lexicale sont représentés par des régions dans un espace sémantique unique muni d'une distance, et le processus de détermination du sens en contexte est modélisé par un système dynamique opérant sur cet espace. Comme nous le verrons, l'utilisation de ressources textuelles électroniques (dictionnaire de synonymes et très gros corpus) permet d'implémenter ce modèle et donc de réaliser un outil de désambiguïsation opérationnel.

2. Les différentes formes de polysémie

2.1. *Polysémie et homonymie : continuités et ruptures de sens*

Deux caractéristiques permettent de définir la polysémie ([KLE 99], p. 55) :

(i) une pluralité de sens liée à une seule forme

(ii) des sens qui ne paraissent pas totalement disjoints, mais se trouvent unis par tel ou tel rapport.

Cette définition s'oppose à celle de l'homonymie où seule (i) est valable. Ainsi dira-t-on que *plateau* est polysémique parce qu'il prend des sens différents dans *un plateau à fromages* et *un plateau de théâtre*, et que ces sens sont unis par l'évocation commune d'une forme horizontale sur laquelle peuvent être disposées un certain nombre de choses. En revanche, on parlera d'homonymie dans le cas de *canon* parce que ses deux sens dans *un tir au canon* et *le droit canon* sont totalement disjoints. Alors que *plateau* apparaît bien comme une unité lexicale unique malgré sa pluralité de sens, on aura tendance à considérer dans le cas de *canon* que l'on a

2. Sans parler d'usages plus difficiles à classer, parce que jouant sur plusieurs sens à la fois, comme dans *Ces pétroliers qui roulent sur l'or noir* (lu sur www.rtl.fr).

affaire à deux unités distinctes, $canon_1$ et $canon_2$, qui partagent une même forme. Il faut noter cependant que la frontière entre polysémie et homonymie n'est pas très nette, comme en témoignent le caractère assez vague de la formulation de la deuxième caractéristique (ii) et les divergences des dictionnaires sur le nombre d'entrées consacrées à telle ou telle unité. Ainsi, faut-il traiter *table* comme une seule unité polysémique, ou doit-on distinguer un $table_2$ à l'œuvre dans *table de multiplication* du $table_1$ qu'on trouve dans *table de cuisine* ?

Plutôt que de chercher à trancher de manière inévitablement arbitraire, il vaut sans doute mieux accepter l'existence d'un continuum et adopter une définition de la polysémie qui en tienne compte [VIC 96]. Partant du fait que le sens d'une unité lexicale dans une occurrence donnée dépend en partie de ce qu'apporte cette unité elle-même de constant, quel que soit le contexte, et en partie de ce qui est fonction du contexte, on peut classer les expressions suivant l'importance relative de ces deux facteurs. A un extrême, le contexte ne joue aucun rôle : l'expression est monosémique ; son sens est le même dans tous les énoncés, ce sens étant donc entièrement défini par l'apport propre de l'unité (exemples : *tournevis*, *hectolitre*…). A l'autre extrême on trouve les homonymes « purs », dont l'apport constant, commun à tous les emplois, est effectivement nul, puisque le sens peut changer radicalement suivant les énoncés (exemples : *avocat*, *sol*, …). Entre ces deux extrêmes, se trouve le cas général de la polysémie, avec des cas qui tendent vers la monosémie, quand le contexte ne joue qu'un rôle minime (tous les sens recensables sont très proches les uns des autres), et d'autres vers l'homonymie, quand l'apport propre constant est très faible.

Prenons quelques exemples pour illustrer ce dernier point. Soit le mot *bureau*. Il possède quatre sens principaux : un meuble (ex. : *s'asseoir à son bureau*), une pièce (ex. : *ouvrir la fenêtre de son bureau*), un établissement (ex. : *le bureau de poste*, *le bureau de tabac*, etc.), une institution (ex. : *le bureau de l'Assemblée*, *le bureau de l'association*, etc.). Ces différents sens sont indéniablement reliés (comme on le verra ci-dessous, ils sont dans des rapports de *métonymie*), ce qui signifie que l'on a bien affaire à de la polysémie et non de l'homonymie. Cependant, quand on essaie de déterminer l'apport propre du mot *bureau* qui est commun à tous ses emplois, on s'aperçoit qu'il est très ténu : une vague notion d'activité d'écriture, qu'il semble difficile de rendre opératoire dans un calcul effectif du sens de *bureau* en contexte. Autrement dit, on se trouve plus près du pôle homonymique dans le continuum qui va de la monosémie à l'homonymie.

A l'inverse, prenons le mot *livre*. Il a aussi plusieurs sens : il peut désigner notamment un objet physique (*un petit livre, de couverture rouge, posé sur l'étagère*), une production intellectuelle (*un livre très drôle, mais très mal écrit*), un produit commercial (*un livre trop cher et introuvable, bien que paru récemment*). Mais l'apport propre de ce mot, commun à ces différents empois, est ici beaucoup plus important. Il est clair qu'un processus de calcul du sens peut utiliser avec profit une notion « générique » de livre, en tant qu'entité abstraite, possédant un contenu intellectuel, produit à de multiples exemplaires pouvant être vendus. Cette notion générique peut ensuite être spécifiée pour aboutir à un sens précis dans un contexte particulier. On a donc affaire dans ce cas à une polysémie beaucoup plus proche de la monosémie que de l'homonymie.

Il faut noter que la situation est généralement plus complexe. Prenons l'exemple du mot *pied*. On peut trouver dans un grand nombre de ses emplois un apport de ce mot que l'on peut formuler de la manière suivante : il désigne la partie la plus basse d'un objet physique, en contact avec une surface horizontale[3]. Cela s'applique bien sûr à l'extrémité du corps humain que l'on appelle ainsi, mais aussi au pied de la table, de l'arbre, du verre, de la falaise, etc. Et cela semble d'autant plus important de caractériser cet apport propre qu'il est très productif : il existe des centaines de noms N pour lesquels le sens de l'expression *pied de N* peut être calculé en utilisant cette formulation. Mais il y a aussi d'autres sens du mot *pied* pour lesquels cette notion de partie basse n'est pas pertinente, comme le sens d'unité de mesure (*haut de six pieds*) ou celle d'unité rythmique (*un vers de six pieds*)[4]. On a donc intérêt, pour des cas comme celui-ci (et ils sont nombreux), à concevoir des modélisations à plusieurs niveaux : un niveau, proche de l'homonymie, qui opère une première séparation entre des sens trop éloignés pour que la caractérisation d'un apport commun soit utilisable, et un deuxième niveau, où l'on peut, pour chaque sous-ensemble de sens, rendre opérationnelles des formulations de l'apport propre qui facilitent le calcul du sens exact de l'unité dans un énoncé donné.

2.2. *Phénomènes de parole et degrés de lexicalisation*

La polysémie est un phénomène « vivant », au sens où les mots sont en permanence susceptibles d'acquérir de nouveaux sens (et aussi d'en perdre d'autres) : c'est l'un des moteurs les plus importants de l'évolution du lexique d'une langue. Deux grands procédés sont principalement à l'origine de l'acquisition de ces nouveaux sens : la *métaphore* et la *métonymie*. La métaphore consiste à utiliser un mot qui désigne habituellement une entité ou un événement d'un certain domaine pour évoquer une entité ou un événement qui joue un rôle analogue dans un autre domaine. Par exemple, c'est par métaphore que l'on parle de *virus informatique* : le mot *virus*, qui vient du domaine de la biologie, est utilisé pour parler de programmes informatiques dont le comportement rappelle celui des virus biologiques. Quant à la métonymie, c'est le procédé qui consiste à évoquer une entité (ou un événement) par le mot qui désigne une autre entité (ou événement), liée à la première par un rapport fonctionnel ou structurel. Ainsi c'est par métonymie que *le premier violon* désigne un violoniste, ou que l'on dit *faire rire la salle* alors que ce sont les occupants de la salle qui rient. Pour reprendre les exemples examinés ci-dessus, c'est par métonymies successives que le mot *bureau* a acquis ses différents sens au cours de l'évolution du français, passant du meuble à la pièce contenant ce meuble, puis au lieu constitué de telles pièces, puis aux groupes travaillant dans de tels lieux[5], alors que c'est par métaphore que le pied, partie de corps humain, a pu aussi désigner la partie correspondante d'une table ou d'un arbre[6].

---

3. Il ne faut pas prendre cet « apport » du mot pour une sorte de définition. En l'occurrence, notre formulation n'est pas une définition de *pied* parce que bien des parties basses en contact avec le sol ne peuvent pas être désignées par *pied*, à commencer par les pattes des animaux.

4. Sans parler de locutions figées, comme *prendre son pied*, *au pied de la lettre* ou *faire le pied de grue*, qui, comme nous l'avons vu (chap. 2, § 1.1) peuvent être traitées comme des unités à part entière.

5. Dans son premier sens, perdu aujourd'hui, il désignait une étoffe (celle de la bure des moines) que l'on plaçait sur les tables sur lesquelles on lisait et écrivait : le sens de meuble est donc déjà le résultat d'une première métonymie.

6. En revanche, c'est par métonymie qu'il a acquis son sens d'unité de mesure (longueur évaluable avec son pied).

La métaphore permet de produire des sens nouveaux de manière quasiment illimitée au cours du discours : on parle de « métaphore vive » [RIC 75]. La métaphore vive est une création éphémère de la parole. C'est « un rapprochement soudain entre des choses qui semblaient éloignées » ([RIC 75], p. 49) :

Son bureau est un hall de gare

La structure du chromosome est tout à la fois code législatif et pouvoir exécutif

Il en est de même pour la métonymie. Ainsi, on caractérise comme des *métonymies vives* les emplois suivants [NUN 78], [FAU 84] :

L'omelette au jambon est parti sans payer ! (énoncé par un serveur de restaurant s'adressant au cuisinier)

L'appendicite du troisième a encore fait de la fièvre cette nuit. (une infirmière à un médecin à l'hôpital).

Mais un certain nombre de ces emplois « passent dans la langue », progressivement, au sens où ils deviennent d'emploi banal, au point d'être répertoriés dans le dictionnaire. On parle alors de métaphores ou de métonymies *lexicalisées* ou *conventionnelles*. Ainsi, dans *J'ai une montagne de choses à faire*, le sens métaphorique de *montagne* n'est plus ressenti comme une figure de rhétorique, mais comme un sens de plein droit du mot *montagne*. De même pour l'emploi métonymique de *bouteille* dans *boire une bonne bouteille.* Ce processus de lexicalisation peut aller jusqu'à la perte totale, dans la conscience des locuteurs, de l'existence même d'un trope à l'origine du sens dérivé. On parle alors de métaphore ou de métonymie « morte ». Ainsi *voler* au sens de dérober et *voler* au sens de se déplacer dans les airs sont considérés aujourd'hui comme des homonymes, le rapport de sens entre eux s'étant complètement perdu, même si, en fait, le premier dérive du second par un processus métaphorique (on disait en français classique : *Le faucon vole sa proie*).

Là encore, il existe bien des cas intermédiaires, entre métaphore (ou métonymie) vive et lexicalisée, ou lexicalisée et morte, qui sont le reflet en synchronie de l'état plus ou moins avancé des processus diachroniques qui tendent à banaliser de plus en plus un certain nombre d'inventions discursives des locuteurs, jusqu'à ce que leur origine devienne parfaitement opaque.

Peut-on calculer des sens métaphoriques ou métonymiques ? Le problème se pose de manière particulièrement aiguë pour les sens non lexicalisés, qui, par définition, ne sont pas recensés dans les dictionnaires. En effet, la tâche est dans ce cas beaucoup plus ardue : le calcul du sens ne consiste pas simplement à sélectionner un sens pertinent parmi un ensemble de sens déjà répertoriés ; il faut « découvrir » le nouveau sens produit par la métaphore ou la métonymie vive. Ainsi Duvigneau [DUV 02], [DUV 03], a trouvé dans un texte de physique de Poincaré l'exemple suivant : *Faudra-t-il chercher à raccommoder les principes ébréchés (...) ?* Les sens de *raccommoder* et de *ébrécher* dans ce contexte n'ont aucune chance de se trouver dans le lexique. Duvigneau analyse ce type de métaphores verbales comme un processus de co-hyponymie (le terme métaphorique *raccommoder* et les termes plus conventionnels *réviser*, *remanier* étant des hyponymes d'un même hyperonyme *réparer*). On a donc dans ce cas une

méthode de calcul du sens métaphorique, qui semble implémentable si l'on dispose d'une ressource lexicale fournissant les relations d'hyperonymie[7].

Malheureusement, toutes les métonymies et les métaphores vives ne sont pas susceptibles d'un tel traitement. Notamment, les exemples que nous avons donné au début de cette section, qu'il s'agisse du *code législatif* de Ricœur ou de *l'omelette au jambon* de Fauconnier, montrent que l'on a souvent besoin d'un contexte très large (incluant la situation d'énonciation) et de connaissances précises sur le monde pour pouvoir interpréter correctement ces expressions. Il faut donc admettre que l'on se trouve là au delà de ce dont on est capable en traitement automatique aujourd'hui et vraisemblablement pour encore de longues années…

2.3. *Métonymies intégrées et coercition de type*

Un certain nombre de phénomènes de changement de sens sont cependant suffisamment systématiques pour pouvoir être traités par des règles générales. C'est le cas notamment de ce que Kleiber [KLE 94] [KLE 99] a appelé les *métonymies intégrées*. Prenons les exemples suivants :

Je suis garé sur la place du marché
George Sand est sur l'étagère de gauche
Paris a massivement voté « oui » au référendum

Il est clair que ce n'est pas « moi » qui suis garé sur la place, mais ma voiture. De même ce n'est pas George Sand mais un exemplaire d'un livre dont elle est l'auteur qui a été rangé sur l'étagère, et ce n'est pas Paris qui a participé au référendum mais bien ses habitants. Dans le premier exemple, on peut d'ailleurs remplacer *je* par *mon frère*, *l'un des suspects* ou *l'épicier du coin*, ce serait toujours d'un véhicule qu'il s'agirait, ce qui montre que ces phénomènes sont vraiment systématiques : c'est toute la classe des groupes nominaux désignant des humains qui possède la capacité de désigner un véhicule dans la position sujet de *être garé* et de bien d'autres prédicats (*rouler*, *déraper*, *entrer en collision*, etc.). Il serait bien sûr complètement inefficace de vouloir conserver dans le lexique cette information pour toutes les unités lexicales concernées. Il faut au contraire entrer ces règles générales en tant que telles avec des mécanismes de déclenchement de type « résolution de conflits » (cf. chap. 5, où de tels mécanismes sont utilisés pour le calcul du temps et de l'aspect)[8].

Pour traiter ces phénomènes, plusieurs auteurs ont cherché à définir des mécanismes généraux qui puissent rendre compte de ces changements systématiques de sens. On peut ainsi citer Pustejovsky [PUS 95] qui propose un mécanisme de *coercition de type*, Nunberg et Zaenen [NUN 97] qui défendent l'idée de *polysémie systématique* (cf. aussi [NUN 95]), ou encore la notion, plus complexe, de *facettes*, défendue par Cruse (cf. [CRU 00] et [CRO 04] chap. 5). Il faut cependant éviter les généralisations trop hâtives et distinguer soigneusement ce qui relève du discursif et qui est effectivement systématique (comme les exemples que nous avons donnés ci-

7. Cf. le travail de Gaume [GAU 02], [GAU 03], qui a conçu une mesure sur des graphes lexicaux, la *proxémie*, qui permet, entre autres, d'obtenir automatiquement ces co-hyponymes.
8. On trouvera au chap. 8 (§§ 2.5 et 2.6) un exemple de méthode de résolution des métonymies intégrées conducteur/véhicule (qui fonctionne d'ailleurs dans les deux sens) dans un système d'Extraction d'Information.

dessus) de ce qui relève de phénomènes lexicaux, qui leur ressemblent à première vue mais qui s'avèrent beaucoup plus rétifs à la généralisation parce que moins systématiques. Il en est ainsi, par exemple, de la *fonction de transfert* postulée par Nunberg et Zaenen [NUN 97] qui régulerait l'emploi d'un nom comptable dans un sens massif. Le passage de comptable à massif ferait appel à un « broyeur universel »[9] : *un lapin*, en devenant *du lapin*, est transformé en « substance lapine » qui peut selon le contexte désigner de la viande de lapin (*J'ai mangé du lapin*), de la fourrure de lapin (*Elle porte du lapin*), ou un mélange indifférencié résultant d'un broyage effectif (*Après que plusieurs camions eurent roulé sur le corps, il y avait du lapin partout sur l'autoroute*). En fait, Kleiber ([KLE 99], chap. 4) montre qu'il faut distinguer le dernier exemple (le lapin sur l'autoroute) des deux précédents (le lapin cuisiné et le lapin en manteau). Seul le lapin sur l'autoroute est effectivement le résultat d'un processus systématique (qui porte bien son nom de « broyeur »), applicable à n'importe quelle entité matérielle, mais dans des conditions discursives très contraintes (il faut une situation très particulière, ici la route et les camions, pour que ce sens soit évoqué[10]). En revanche, les acceptations 'viande de lapin' et 'fourrure de lapin' doivent être considérées comme lexicalisées, faisant partie du potentiel sémantique de l'unité lexicale *lapin*, et n'étant pas inférable par une règle générale. En effet, si *de la mirabelle* désigne de l'alcool de mirabelle, *du raisin* ne peut pas dénoter de l'alcool de raisin ou du vin, de même que *de l'orange* n'est pas du jus d'orange, *de l'olive* n'est pas de l'huile d'olive, etc. L'exemple du mot *vison* montre bien d'ailleurs que ces processus sont spécifiques de l'unité considérée : si *du vison* désigne bien de la fourrure de vison, *un vison* dénote plus facilement un manteau qu'un animal, alors que l'hypothèse de fonctions de transfert devrait en faire un sens doublement dérivé (un premier transfert de l'animal « broyé » en fourrure, puis un deuxième transfert, en sens opposé massif → comptable, « découpant » un vêtement dans ladite fourrure…).

D'une manière générale, il faut donc distinguer deux niveaux de modélisation, qui correspondent à deux « temps » dans le processus de calcul du sens. D'abord, on doit modéliser un processus essentiellement lexical, dans lequel on s'appuie sur les sens conventionnels des unités polysémiques, comme par exemple les sens 'animal' et 'fourrure' pour *vison*, 'fruit' et 'liqueur' pour *mirabelle*, 'animal' et 'viande' pour *poulet*, 'animal', 'viande' et 'fourrure' pour *lapin*[11], pour sélectionner le sens approprié. Et ce n'est que dans un deuxième temps que les règles discursives doivent être mises en œuvre, si les contraintes imposées par le contexte phrastique l'imposent. Ainsi *du lapin* que l'on mange ou que l'on porte sera analysé comme tel sur une base lexicale, et ce n'est que dans le cas de l'autoroute que le sens 'animal' (dans la mesure où il aura été sélectionné dans le processus précédent), de type comptable, provoquant un conflit avec le partitif *du* qui exige le type massif, conduira à l'utilisation de mécanismes discursifs de changement de type tel que le broyeur universel pour aboutir à l'interprétation pertinente.

9. L'expression *universal grinder* a été introduite par Pelletier [PEL 75], selon [KLE 99].

10. Le broyage n'est pas le seul mécanisme qui peut opérer la conversion comptable → massif au niveau discursif. Dans *Il y a du sanglier dans cette forêt*, il s'agit, comme le rappelle Kleiber, de sangliers bien entiers ! Citons aussi *Ça, c'est de la belle armoire !* qui n'implique pas, loin de là, que l'armoire en question en morceaux.

11. Il faut noter que la forme sous laquelle ces sens sont consignés dans le lexique peut être très variée. Notamment, on peut utiliser les principes du lexique génératif [PUS 95] pour introduire, à ce niveau, le maximum de systématicité.

2.4. *Représentations de la polysémie lexicale*

Si l'on en vient maintenant à l'information qu'il faut attacher en propre à chaque unité lexicale, dont on vient de voir qu'elle constitue, malgré l'existence de mécanismes généraux, une part irréductible et essentielle pour la plupart des unités polysémiques, on peut se demander sous quelle forme cette information doit être structurée.

En gros, deux hypothèses différentes (cf. fig. 1) ont été proposées par les différentes théories sémantiques lexicales qui cherchent à préserver l'unicité de l'unité lexicale face à la prolifération des sens :

- *Hypothèse 1 :* on postule l'existence d'un *noyau de sens* commun, ou potentiel de sens[12], à partir duquel s'opère une « dérivation » des divers sens. Divers mécanismes peuvent alors jouer pour spécifier un sens particulier à partir du noyau commun : par spécialisation (le noyau de sens étant alors vu comme générique) et/ou par enrichissement, par apport d'un contexte d'usage.

- *Hypothèse 2 :* on postule l'existence d'un *sens premier, de base*[13], dont les autres sens seraient dérivés par des opérations (généralement métaphores et métonymies) aboutissant à des sens conventionnels différents, souvent relatifs à différents « domaines » (de pensée, de connaissance…).

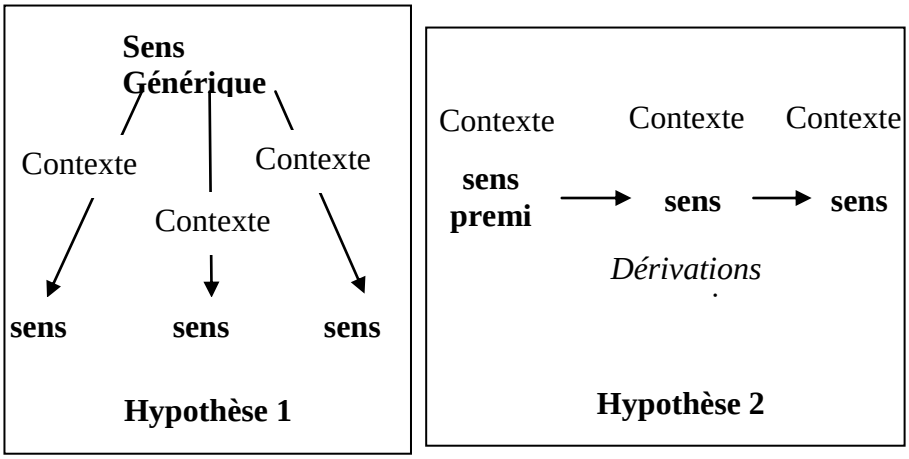


Figure 1. *Les deux hypothèses de structuration de la polysémie lexicale*

Pour illustrer concrètement ces deux hypothèses, reprenons l'exemple du verbe *rouler* (cf. chap.2, § 1.3). Voici une liste (non exhaustive !) d'emplois de *rouler*

(1) *rouler un tonneau*
(2) *rouler un tap*is
(3) *rouler une cigarette*
(4) *rouler dans un véhicule*
(5) *rouler sur soi-même*

12. La terminologie est très variable suivant les auteurs et les courants de pensée : les *formes schématiques* de Culioli [CUL 90], les *archétypes cognitifs* de Desclés [DES 85], les *motifs* de Cadiot et Visetti [CAD 01] ou encore les *schémas* de Langacker [LAN 87] sont autant de théorisations, bien différentes les unes des autres, de cette hypothèse générale.

13. Là aussi la terminologie est très variable : pour ne donner qu'un exemple Picoche, utilisant un vocabulaire guillaumien, parle de *saisie plénière* [PIC 86].

(6) *se rouler en boule*
(7) *rouler les hanches*
(8) *le bateau roule et tangue*
(9) *rouler les r*
(10) *rouler quelqu'un*
(11) *rouler pour quelqu'un*
(12) *ça roule !*
(13) *rouler sa bosse*

Dans le cadre de l'hypothèse 1, le noyau de sens pourrait être la notion de rotation, déclinée de différentes manières et enrichie d'éléments variés dans les emplois de (1) à (8). Ainsi le complément d'objet spécifierait le but de la rotation dans les exemples (1)-(3) : on roule un tonneau pour le déplacer, un tapis pour le ranger, une cigarette pour la fabriquer etc. Mais les emplois de (9) à (12) doivent être alors exclus et traités indépendamment puisqu'ils ne contiennent pas du tout l'idée de rotation, sans parler de la locution figée (13) qui doit, de toute façon, être traitée comme telle (cf. chap. 2, § 1.1). On peut sans doute « récupérer » un emploi tel que (9) en affaiblissant quelque peu ce noyau si l'on ne retient que l'idée de mouvement périodique, mais cela ne résout pas pour autant le cas des emplois plus abstraits de (10)-(12), qui sont, il est important de le souligner, extrêmement fréquents dans les corpus généralistes et le langage courant…

Une autre solution, sans doute plus satisfaisante pour la plupart des théories qui défendent l'hypothèse 1, serait de construire un noyau de sens à partir de la description même des emplois les plus « physiques » du verbe, mais en en extrayant une caractérisation plus abstraite. Si l'on oppose *rouler* à d'autres verbes exprimant un déplacement, comme *glisser*, *marcher*, *voler*, etc., on remarque que *rouler* se caractérise, du point de vue de la mécanique, par l'utilisation des forces de frottement, qui s'opposent a priori au déplacement, pour avancer quand même, sans perte d'énergie. C'est bien sûr la roue qui réalise le mieux ce type de mouvement, mais le même principe est à l'œuvre quand on roule un bloc de pierre : c'est parce que la force de frottement s'oppose au mouvement de translation (le glissement) que l'on peut basculer le bloc pour le déplacer dans la direction voulue. On peut donc en tirer, par abstraction, un noyau de sens comme 'utiliser une force de résistance pour parvenir à ses fins' qui aurait l'avantage de s'appliquer aussi bien au sens physique de déplacement qu'au sens à l'œuvre dans l'exemple (10) : qu'est-ce donc en effet que *rouler quelqu'un* sinon de le conduire à une situation qu'il ne désirait pas en utilisant ses réticences plutôt qu'en s'y opposant directement. Mais, même si cette solution présente l'avantage d'éviter la coupure entre sens « littéral » et sens « figurés », elle échoue à rendre compte de certains emplois comme (6), (7) ou (11), et elle est à l'évidence inutilisable par un système automatique.

Passons maintenant à l'hypothèse 2. On choisira vraisemblablement comme sens premier le sens de 'déplacement par un mouvement de rotation', à l'œuvre dans les emplois (1) et (5), qui est d'ailleurs son sens étymologique[14]. On obtient ensuite les autres sens par des opérations de métonymie et de métaphore. Ainsi on dit qu'un véhicule roule par métonymie (ce sont ses roues

14. Les deux seuls emplois attestés au 12è siècle sont, selon le *Trésor de la langue française*, « effectuer un mouvement en tournant sur soi-même » et « déplacer un objet en le faisant tourner sur lui-même ».

qui roulent), et une deuxième métonymie, entre le conducteur et son véhicule, conduit au sens (4). Le sens (2) serait, lui, un sens métaphorique : les gestes effectués pour rouler un tapis sont semblables à ceux que l'on fait pour rouler un tonneau, même si le résultat obtenu n'est pas le même. Et on pourrait déduire le sens (3) du sens (2), grâce à une métonymie supplémentaire : c'est la feuille de cigarette qu'on roule comme on roule un tapis. On peut sans doute aussi avancer que les sens (6), (7), et peut-être (8) sont aussi le fruit de métaphores, à partir de l'image produite par le mouvement de rotation d'un corps sur lui-même. Mais comme on s'en rend bien compte, l'exercice atteint là aussi assez vite ses limites, et, quelle que soit l'ingéniosité avec laquelle on peut rattacher les sens (9) à (12) par des procédés associatifs plus abstraits, on aboutit à la même conclusion que pour l'hypothèse 1 : ces procédés deviennent inutilisables par des processus automatiques, parce que l'on serait bien en peine de les formaliser par des règles générales suffisamment précises. Tout juste peut-on garder l'idée que l'ensemble de ces sens peut être décrit comme une « ressemblance de famille » [WIT 53], [LAK 87] dans le cadre théorique que Kleiber [KLE 90] appelle « la théorie du prototype étendue ».

On doit donc en conclure qu'il ne sert à rien, pour les besoins du traitement automatique, de chercher à tout prix à maintenir l'unicité de représentation des unités lexicales polysémiques. On doit adopter une attitude beaucoup plus pragmatique, et s'adapter au types de tâche et de corpus auxquels on a affaire. Comme on l'a dit en introduction, pour certains types de tâches et de corpus, on peut même en arriver à faire l'impasse totale sur la polysémie de certaines unités. Ainsi, dans une tâche d'extraction d'information sur un corpus de constats d'accidents de la route (cf. chap. 8), limiter la représentation du verbe *rouler* au sens (4) n'entraînera vraisemblablement que très peu d'erreurs. Dans des corpus moins spécialisés, si la tâche ne réclame pas une grande finesse dans la détermination du sens du verbe, on pourra aussi se contenter d'un découpage très grossier, en regroupant les différents sens dans des grandes rubriques, telles que 'déplacement' pour (1), (4) et (5), 'changement de forme' pour (2), (3) et (6), 'tromperie' (regroupant 'vol', 'mensonge', etc.) pour (10), et ainsi de suite.

Mais si l'on veut modéliser de façon plus complète et plus précise le comportement sémantique de l'unité polysémique considérée, on est rapidement confronté à de nouvelles difficultés : combien doit-on prévoir de sens différents pour un verbe comme *rouler* ? A quel degré de finesse doit-on s'arrêter ? Et comment sélectionner « le » sens pertinent dans un énoncé si l'on multiplie des représentations à la fois concurrentes et très proches ? Comme nous le discuterons en détail sur des exemples ci-dessous, ce dernier problème se pose de manière d'autant plus aiguë que, suivant les énoncés, une unité peut avoir un sens plus ou moins précis, voire de relever de plusieurs sens à la fois. Ces difficultés sont bien sûr le « prix » payé pour l'abandon d'une représentation sémantique unifiée pour chaque unité polysémique. Nous allons les aborder ici dans le cadre d'un modèle particulier de la polysémie, dont la caractéristique essentielle est de proposer une représentation continue des variations de sens d'une unité polysémique. Comme nous le verrons, cela permet de rendre compte de l'aspect graduel de ces variations, et donc de pouvoir travailler à différents degrés de finesse en fonction des besoins et des énoncés.

3. Une représentation géométrique de la polysémie

Nous avons proposé, il y a quelques années [VIC 96], un modèle mathématique de la polysémie, qui fait correspondre à chaque unité polysémique un *espace sémantique*, dans lequel sont représentés l'ensemble de ces sens. Nous allons montrer ici comment cet espace peut être construit automatiquement en utilisant un graphe de la relation de synonymie, et, dans un deuxième temps, comment l'on peut calculer le sens pertinent de l'unité dans un énoncé donné. Pour illustrer concrètement nos méthodes, nous prendrons l'exemple de l'adjectif *sec*[15].

3.1. *Représentation des différents sens d'une unité polysémique*

L'adjectif *sec* est très polysémique : le *Trésor de la Langue Française* en donne plus de trente acceptions. En voici les principales :

1. qui ne contient pas d'eau. Ex.: *La route était sèche*.
2. maigre, décharné. Ex.: *Un vieil homme sec et ridé*.
3. stérile, improductif. Ex.: *Rester sec à un examen*.
4. insensible, sévère, égoïste. Ex.: *Un homme au cœur sec*.
5. brusque, abrupt. Ex.: *Donner un coup sec*.
6. simple, seul. Ex. *Avoir un atout sec dans son jeu*.

Bien que ces sens soient nettement différents, ils entretiennent un certain nombre de relations de voisinage. Ainsi les sens 1, 2 et 3 sont concomitants quand *sec* qualifie de la végétation. Le lien entre les sens 3 et 4 est aussi clair : une personne égoïste est quelqu'un qui ne donne pas beaucoup d'elle-même. Le sens 5, qui s'applique à des événements, est proche du sens 4 quand il caractérise un comportement brutal. Enfin le sens 6 peut être relié aussi bien au sens 5 qu'aux sens 2 et 3 : un atout sec est une carte qui manque « d'accompagnement » pour être pleinement profitable. Ainsi l'ensemble des sens de *sec* peut bien être décrit comme une « ressemblance de famille » avec six sens prototypiques qui se recouvrent en partie. Il est important de noter que chacun de ces sens prototypiques de *sec* rassemble lui-même de nombreuses nuances de sens. Par exemple, on peut distinguer sous le sens 4 différents traits psychologiques : égoïsme, froideur, sévérité, brutalité, etc. Il faut aussi remarquer que, bien souvent, un emploi de *sec* « joue » sur plusieurs sens à la fois. Ainsi *un ton sec* caractérise un ton à la fois du point de vue physique (sens 5) et psychologique (sens 4). De même *une terre sèche* dénote une terre qui manque d'eau (sens 1) et de fertilité (sens 3). Ces cas, que nous appelons *cas d'indétermination*, sont très fréquents, et trop souvent négligés dans les analyses sémantiques. Il faut les distinguer des vrais *cas d'ambiguïté-alternative* dans lesquels deux sens sont aussi possibles, mais de manière exclusive : seul l'un des sens est pertinent dans un énoncé donné. Par exemple, *un homme sec* peut être utilisé pour parler soit d'un homme physiquement maigre soit d'un homme psychologiquement dur, mais très rarement les deux à la fois.

Comme nous l'avons dit, notre modèle associe à toute unité polysémique un espace sémantique, multidimensionnel, qui doit pouvoir représenter tous les sens de l'unité étudiée. Ainsi, pour *sec*, chaque sens prototypique occupera une région de l'espace dans laquelle chaque

15. Ce travail sur l'adjectif *sec* a été mené en collaboration avec Jacques François et Jean-Luc Manguin du laboratoire CRISCO de Caen, et a déjà donné lieu à plusieurs publications : [VEN 02], [VEN 04], [FRA 03].

nuance de sens correspond à un point, et la distance définie sur cette espace doit représenter les différents liens de proximité sémantique entre les sens. Le calcul du sens dans un énoncé donné est modélisé par un système dynamique : les autres unités présentes dans l'énoncé définissent une fonction potentielle sur l'espace sémantique et les valeurs de la fonction inférieures à un seuil donné définissent la région de l'espace correspondant au sens de l'unité étudiée dans l'énoncé en question[16].

Ces principes sont illustrés par le schéma de la figure 2. Le plan dessiné dans le bas de la figure représente l'espace sémantique d'une unité polysémique. Autrement dit, tout point de ce plan correspond à un sens précis de l'unité, et deux points sont d'autant plus proches dans le plan que les sens correspondants sont dans une relation de proximité sémantique, du type de celles que nous avons mis en évidence entre les différents sens de *sec*. Il faut garder à l'esprit que cet espace sémantique peut avoir un grand nombre de dimensions (on s'est limité sur la figure à deux dimensions pour des raisons évidentes de lisibilité), et donc que beaucoup de relations de proximité peuvent être à l'œuvre en même temps. Supposons maintenant que l'on rencontre cette unité polysémique dans un énoncé. Le sens de l'unité dans cet énoncé dépend des autres unités présentes dans l'énoncé ainsi que les relations syntaxiques qu'elles entretiennent avec l'unité en question, c'est-à-dire de ce que l'on appelle les *éléments co-textuels*. Autrement dit, à cause de la présence de ces éléments, certains sens de l'unité seront plus « plausibles », plus « pertinents » que d'autres. C'est ce qu'indique la fonction potentielle dessinée au dessus du plan, et dont on a tracé aussi quelques lignes de niveau sur le plan lui-même. Plus la valeur de cette fonction est basse en un point du plan, plus le sens correspondant est pertinent dans cet énoncé. La fonction que l'on a choisi de dessiner présente deux bassins, centrés respectivement sur les points A et B (dans la théorie des systèmes dynamiques, on appelle ces points les *attracteurs* de la dynamique induite par cette fonction potentielle). Le modèle consiste à considérer que le sens de l'unité dans l'énoncé est donné par l'ensemble des valeurs suffisamment basses de la fonction (inférieures à un seuil donné). Dans notre cas, le sens sera donc représenté par deux régions : une région assez large autour du point A et une région beaucoup plus étroite autour du point B. Cela veut dire que l'unité est ambiguë dans cet énoncé : elle peut avoir le sens très précis associé à B, ou un sens, plus indéterminé, couvrant le sens représenté par à A et tous les sens proches de celui-ci.

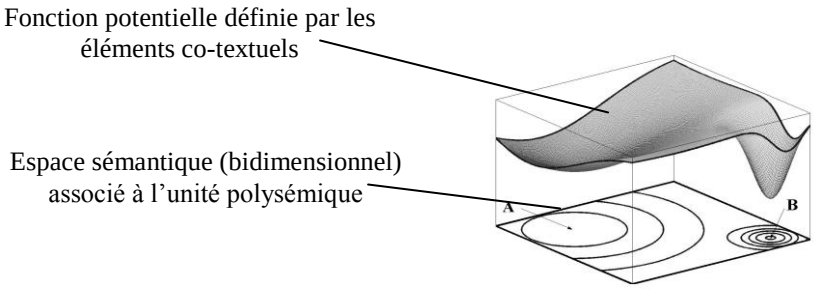
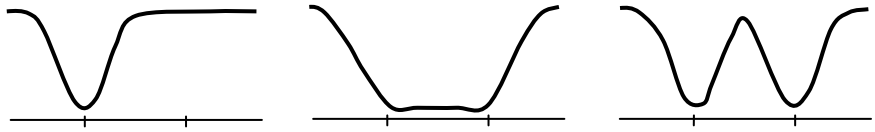


Figure 2. *Représentation d'une fonction potentielle sur un espace sémantique*

16. Pour plus de détails, voir [VIC 96].

On peut ainsi représenter tous les cas de figure interprétatifs dont nous avons parlé à propos de *sec* : sens précis, cas d'indétermination et ambiguïtés-alternatives[17]. On a représenté chacun de ces cas de figure pris isolément sur la figure 3. Cette fois-ci, on a dessiné un espace à une seule dimension, c'est-à-dire un segment de droite. Chacun des cas de figure est illustré par une fonction de forme différente : un bassin très étroit pour le sens précis (a), un bassin très large pour l'indétermination (b), et deux bassins séparés par un col élevé pour l'ambiguïté-alternative (c).



(a) sens (b) (c) ambiguïté-

Figure 3. *Les différents cas de figure interprétatifs (espace unidimensionnel)*

3.2. *Construction de l'espace sémantique à l'aide d'un graphe de synonymie*

Pour implémenter ce modèle, la première étape consiste à construire l'espace sémantique associé à une unité polysémique. Nous avons conçu dans ce but une méthode automatique [PLO 98] qui utilise un dictionnaire électronique de synonymes[18] et un algorithme basé sur les propriétés spécifiques du graphe de synonymie. La relation de synonymie donnée par les lexicographes est une relation de synonymie *partielle* : sont considérés comme synonymes deux mots qui, pour certains contextes, sont remplaçables l'un par l'autre. Généralement ces deux mots ne sont pas remplaçables l'un par l'autre dans tous les contextes. Par exemple, si l'on prend les adjectifs *sec* et *froid*, ils sont considérés comme synonymes parce qu'on peut les remplacer l'un par l'autre, sans trop modifier le sens, dans des expressions telles que *une réponse sèche* ou *un comportement sec*. Mais bien sûr, dans d'autres contextes, la substitution change radicalement le sens : il suffit de comparer, par exemple, *un climat sec* et *un climat froid* ou *un lit sec* et *un lit froid*. Comme beaucoup de mots sont polysémiques, cette relation de synonymie n'est pas transitive : par exemple *sec* est aussi synonyme de *maigre*, alors que *froid* et *maigre* ne le sont pas du tout. En examinant les relations entre les différents synonymes d'un même mot, on peut donc se faire une idée du degré de polysémie de ce mot : c'est cette idée générale que nous avons mise en œuvre dans notre modèle.

17. On trouvera des représentations analogues dans [SAD 86].

18. Ce dictionnaire est consultable en ligne sur le site du CRISCO : http://www.crisco.unicaen.fr. On trouvera aussi sur le site de l'ISC (http://dico.isc.cnrs.fr/), conçu par Sabine Ploux et ses collaborateurs, outre une autre version de ce dictionnaire, un dictionnaire de synonymes de l'anglais, ainsi que des représentations géométriques (« atlas sémantiques »), conçus par Sabine Ploux, qui permettent de visualiser les sens des unités non seulement en contexte monolingue (synonymes) mais aussi en contexte bilingue (traduction) [PLO 03].

Plus précisément, la méthode consiste à calculer les *cliques* du graphe de synonymie contenant l'unité étudiée. Les cliques d'un graphe sont ses sous-graphes complets maximaux[19], c'est-à-dire les ensembles de sommets du graphe qui sont tous reliés entre eux deux à deux. Dans notre cas, chaque sommet est un mot, et deux mots sont reliés entre eux s'ils ont été répertoriés comme synonymes dans le dictionnaire. Une clique est donc un ensemble de mots qui sont tous synonymes deux à deux. On fait l'hypothèse que chaque clique représente un sens très précis (une nuance de sens) de l'unité, que l'on doit donc associer à un point de l'espace sémantique.

Pour pouvoir construire l'espace sémantique lui-même, il faut alors placer ces différents points (représentant chacun une clique) les uns par rapport aux autres. Cela n'est possible qu'à condition de définir une *distance* entre ces points. Cette distance doit permettre de représenter la notion de proximité sémantique qui est au cœur du modèle. Nous avons donc recherché une distance qui réponde à cet objectif.

D'un point de vue technique, cette distance est définie par la métrique du χ^2. Plus précisément, appelons u_1, u_2, …, u_n les synonymes de l'unité, c_1, c_2, …, c_p les cliques associées et posons $x_{ki} = 1$ si $u_i \in c_k$ et $x_{ki} = 0$ si $u_i \notin c_k$. La distance $d(c_k, c_l)$ entre deux cliques est alors donnée par les formules suivantes :

$$d^2(c_k, c_l) = \sum_{i=1}^{n} \frac{x}{x_{\bullet i}} \left(\frac{x_{ki}}{x_{k\bullet}} - \frac{x_{li}}{x_{l\bullet}} \right)^2$$

avec $x_{\bullet i} = \sum_{j=1}^{p} x_{ji}$, $x_{k\bullet} = \sum_{i=1}^{n} x_{ki}$, et $x = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ji}$.

D'un point de vue plus intuitif, il suffit de savoir que cette distance possède les deux caractéristiques suivantes. D'une part, chaque synonyme intervient dans le calcul de la distance entre cliques avec un « poids » d'autant plus faible que le synonyme est présent dans un plus grand nombre de cliques : les synonymes qui sont les moins spécifiques jouent donc un rôle moins important dans la discrimination des sens de l'unité. D'autre part, le nombre d'éléments d'une clique intervient au dénominateur dans les calculs de distance qui concernent cette clique : autrement dit, les cliques composées de beaucoup de synonymes vont être rapprochées les unes des autres, et vont, du coup, avoir tendance à occuper une position plus centrale dans l'espace sémantique.

Pour illustrer concrètement notre méthode, reprenons l'exemple de *sec*. Le dictionnaire que nous utilisons donne 56 synonymes de *sec* :

19. Pour une introduction à la théorie des graphes, voir par exemple [BER 70]. Pour l'algorithme de calcul des cliques, voir [REI 77]. Pour une approche similaire utilisant aussi un graphe de synonymie, voir [WAR 85]. Il faut noter aussi les points communs et les différences avec l'approche de Wordnet [FEL 98] : les *synsets* de WordNet sont analogues aux nos cliques, à la différence près que les synsets sont construits à partir d'une analyse sémantique préalable des mots effectuée par des opérateurs humains (qui ont dû, pour chaque mot, décider du nombre de sens à lui attribuer, et spécifier chacun de ces sens), tandis que pour les cliques, seule la relation de synonymie a été spécifié par des lexicographes.

aigre, âpre, aride, austère, autoritaire, blessant, bourru, bref, brusque, brutal, cassant, concis, cru, décharné, déplaisant, désagréable, désertique, désobligeant, desséché, dur, efflanqué, égoïste, émacié, endurci, étique, étriqué, fauché, ferme, froid, glacé, glacial, improductif, indifférent, ingrat, insensible, maigre, maigrelet, pauvre, racorni, raide, rébarbatif, rebutant, revêche, rigide, rogue, rude, sec, séché, seul, sévère, simple, squelettique, stérile, tranchant, vert, vide.

Rappelons que ce ne sont que des synonymes partiels de *sec* : chaque synonyme ne représente qu'une partie de l'ensemble de ses sens. Le dictionnaire nous donne aussi lesquels sont synonymes entre eux. On a représenté figure 4 une petite partie du graphe de synonymie obtenu.

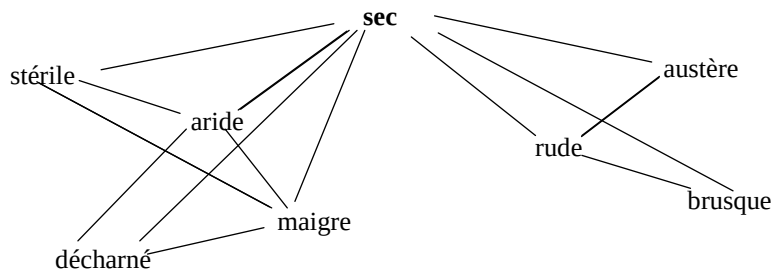


Figure 4 : *Une partie du graphe de synonymie de* sec

Comme on peut constater, un seul synonyme ne suffit généralement pas à définir un sens précis de *sec.* Ainsi, si l'on essaie de se situer par rapport aux cinq principaux sens de *sec* que nous avons listé plus haut (§ 3.1), on constate que *maigre*, par exemple, qui est synonyme à la fois de *décharné* et de *stérile*, couvre les sens 2 et 3. De même, *rude* couvre les sens 4 et 5. Mais les quatre cliques de ce petit graphe définissent des sens plus précis : comme on pourra le vérifier, ce sont {*aride*, *décharné*, *maigre*, *sec*}(sens 2), {*aride*, *maigre*, *sec*, *stérile*}(sens 3), {*austère*, *rude*, *sec*}(sens 4) et {*brusque*, *rude*, *sec*}(sens 5).

En fait, le calcul sur le graphe des 56 synonymes donne 75 cliques, dont voici un échantillon :

{*aride, desséché, décharné, maigre, sec*}, {*aride, désertique, ingrat, sec, stérile*}, {*aride, desséché, maigre, sec, stérile*}, {*aride, improductif, pauvre, sec, stérile, vide*}, {*austère, rude, sec, simple, sévère*}, {*autoritaire, brusque, cassant, sec, tranchant*}, {*bourru, brusque, brutal, cru, rude, sec*}, {*bref, brusque, brutal, sec*}, {*bref, concis, sec*}, {*dur, froid, indifférent, insensible, sec*}, {*fauché, pauvre, sec*}, {*indifférent, insensible, sec, égoïste*}, {*sec, seul, simple*}, …

Comme on peut le constater, chaque clique définit bien une nuance de sens de *sec*. Pour pouvoir étudier l'espace sémantique obtenu par cette méthode, nous avons réalisé un logiciel, VISUSYN, qui permet de visualiser cette espace grâce à des projections bidimensionnelles suivant

les composantes principales du nuage de points constitué par les cliques[20]. La projection sur les deux premières composantes principales est reproduite figure 5.

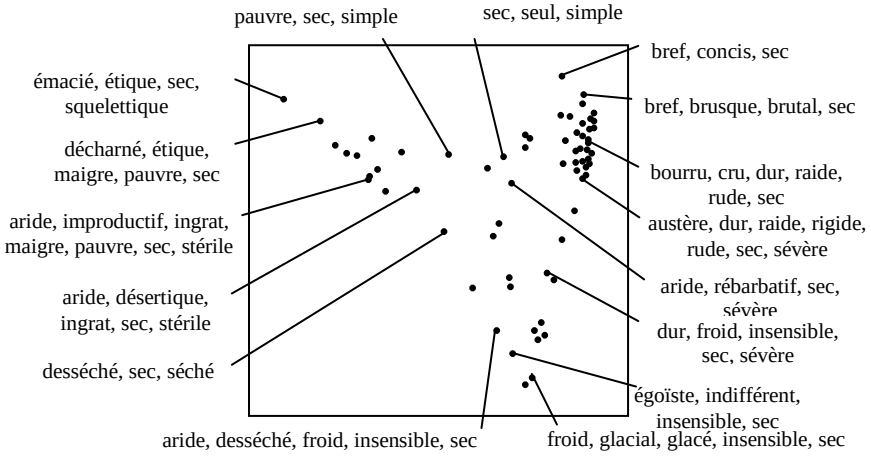


Figure 5: *L'espace sémantique associé à* sec

Les différents sens de *sec* sont clairement identifiables sur la figure. Le sens 1 ('qui ne contient pas d'eau') est situé en plein centre de l'espace (représenté par la clique {*desséché*, *sec*, *séché*}). En allant vers le coin en haut à gauche, on passe progressivement au sens 3 ('improductif') puis au sens 2 ('maigre'). Sur la droite on trouve les sens psychologiques (sens 4), avec en bas les cliques définissant l'égoïsme et l'insensibilité, et, en remontant, la sévérité et la rudesse. Près de cette dernière nuance, on trouve dans le coin en haut à droite le sens 5 (qualifiant des événements). Enfin le sens 6 ('seul') est situé en haut au centre.

On peut vérifier que toutes les relations de proximité sémantique que nous avons décrites plus haut (§ 3.1) sont bien prises en compte : entre les sens 1, 2 et 3 ; entre les sens 3 et 4 ; entre les sens 5 et 4, entre les sens 6 et 5 ; et enfin entre les sens 6, 2 et 3. Cette méthode permet donc effectivement de construire une représentation du sens qui répond à nos objectifs, et cela de manière entièrement automatisée.

4. Calcul du sens d'une unité polysémique dans un énoncé donné

La deuxième étape de notre programme de recherche consiste à obtenir automatiquement le sens d'une unité polysémique, c'est-à-dire, dans notre modèle, à calculer la fonction potentielle pertinente associée à un énoncé donné. Ce travail n'en est qu'à ses débuts. Nous avons mis au point une méthode que nous avons utilisée avec succès sur quelques adjectifs polysémiques, pour lesquels nous ne nous intéressons qu'à l'influence du nom régissant (dont ils sont épithètes) sur leur sens. Nous sommes à l'heure actuelle en train de compléter cette méthode, notamment en

20. Pour d'autres exemples d'application de la méthode à d'autres unités polysémiques, voir [MAN 99], [MAN 01], [MAN 02], [VIC 02], [FRA 05].

utilisant des outils d'analyse syntaxique plus poussés, de manière à pouvoir la généraliser aux unités polysémiques d'autres classes syntaxiques, notamment les verbes. Nous allons donc commencer par décrire la méthode en utilisant toujours l'exemple de *sec*[21], puis nous présenterons les travaux en cours sur les verbes.

4.1. *Principes de la méthode de calcul du sens*

Notre méthode consiste à utiliser un grand corpus[22], dont nous extrayons toutes les occurrences de l'adjectif étudié, ce qui nous permet de relever les principaux noms dont il est épithète. Voici un échantillon de la liste obtenue pour *sec* :

air, arbre, bois, boue, bras, chambre, chemin, cheveu, chose, claquement, cœur, coin, corps, cou, coup, doigt, éclat, esprit, façon, femme, feuille, figure, fleur, foin, fromage, geste, gorge, herbe, homme, jambe, jardin, lèvre, ...

Ensuite nous relevons dans le corpus toutes les occurrences de ces noms avec chacun des synonymes de l'adjectif étudié. A partir de ces données, nous calculons le *degré d'affinité* d'un nom avec une clique adjectivale. Ce degré d'affinité, compris entre 0 et 1, est d'autant plus grand qu'il y a plus d'occurrences dans le corpus du nom avec chaque adjectif de la clique. Plus précisément, appelons w_1, w_2, …, w_m les noms (avec les notations du § 3.2 : u_1, u_2, …, u_n dénotent les synonymes adjectivaux,, c_1, c_2, …, c_p les cliques, et $x_{ki} = 1$ ssi $u_i \in c_k$) et notons n_{ij} le nombre d'occurrences du couple (w_i , u_j) dans le corpus. Ce nombre doit être pondéré par la plus ou moins grande fréquence de w_i et de u_j, pris indépendamment, dans le corpus. S'il n'y avait pas d'affinité particulière entre certains noms et certains adjectifs, les couples seraient équidistribués : autrement dit, le nombre d'occurrences du couple (w_i , u_j) ne devrait être fonction que de la fréquence des deux mots pris indépendamment. Appelons m_{ij} ce nombre moyen « théorique ». On peut montrer facilement que l'on a :

$m_{ij} = \frac{n_{i\bullet}.n_{\bullet j}}{n}$ (avec les mêmes conventions qu'au § 3.2)

Pour mesurer l'affinité d'un nom et d'un adjectif, il faut donc comparer n_{ij} et m_{ij}. Si n_{ij} est nettement supérieur à m_{ij}, cela veut dire que le nom et l'adjectif entretiennent une relation d'affinité particulière. Si au contraire n_{ij} est nul ou nettement inférieur à m_{ij}, cela signifie que le couple est non attesté ou très rare. Nous définissons donc le degré d'affinité d_{ij} du nom w_i avec l'adjectif u_j de la manière suivante :

$d_{ij} = f\left(\frac{m_{ij}}{n_{ij}}\right)$ avec $f(x) = \frac{x}{2}$ ssi $0 < x < 2$ et $f(x) = 1$ ssi $x > 2$

(le degré d'affinité est donc toujours compris entre 0 et 1).

---

21. On trouvera dans [FRA 05] et [MAN 02] des exemples de ces calculs pour d'autres adjectifs (*curieux*, *gros* et *gras*).
22. Il s'agit en l'occurrence du corpus Frantext (http://frantext.inalf.fr/), pour cette étude sur l'adjectif *sec*. Sur l'utilisation de corpus en traitement automatique des langues, voir [HAB 97], [PIE 00], [BOU 01]. Voir aussi [IDE 90] pour une autre approche de désambiguïsation automatique à l'aide d'un dictionnaire électronique.

Pour calculer le degré d'affinité d'un nom avec une clique, on fait alors la somme pondérée des affinités du nom avec toutes les unités qui constituent la clique. Plus précisément, le degré d'affinité a_{ik} du nom w_i avec la clique c_k est donné par la formule suivante :

$$a_{ik} = \frac{\sum_j d_{ij}.p_{ij}.x_{kj}}{\sum_j p_{ij}.x_{kj}}$$ où le facteur de pondération p_{ij} vaut $\frac{m_{ij}}{\sum_k x_{kj}}$

Pour illustrer ce calcul, prenons l'exemple du nom *terre* et du sous-ensemble de synonymes que nous avons utilisé au § 3.2 figure 4. (*aride*, *austère*, *brusque*, *décharné*, *maigre*, *rude*, *sec*, *stérile*) . Les nombres de co-occurrences dans le corpus sont présentés dans le tableau 1 ci-dessous :

terre aride	8	*terre maigre*	2
terre austère	1	*terre rude*	1
terre brusque	0	*terre sèche*	3
terre décharnée	0	*terre stérile*	6

Tableau 1: *Fréquence de co-occurrence de* terre *avec les synonymes*

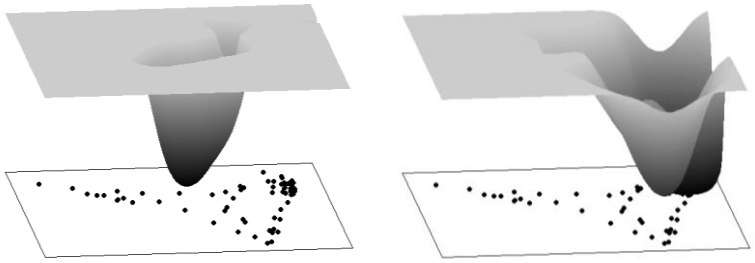
Le nom *terre* est donc bien attesté avec les adjectifs *aride* et *stérile* (représentatifs du sens 3), dans une moindre mesure avec *maigre*, *rude* et *austère*, et pas du tout avec *brusque* et *décharné*. Le calcul des degrés d'affinité avec les cliques rend compte de ces observations, puisque l'on obtient les résultats suivants, présentés dans le tableau 2:

{*sec, aride, maigre, stérile*}	0,98
{*sec, aride, maigre, décharné*}	0,57
{*sec, rude, austère*}	0,42
{*sec, rude, brusque*}	0,28

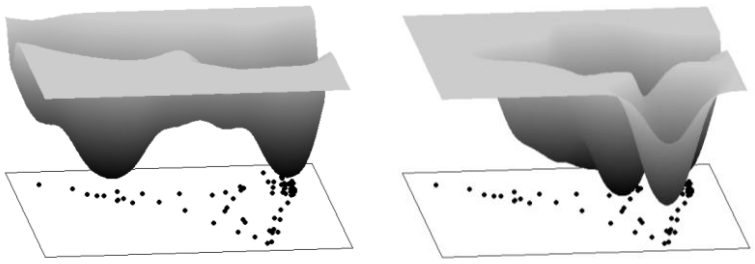
Tableau 2: *Degré d'affinité de* terre *avec les cliques*

On conçoit donc que le degré d'affinité permette de calculer l'influence qu'il exerce sur le sens de l'adjectif. Nous avons implémenté ces calculs dans VISUSYN, de manière à visualiser la fonction potentielle obtenue à partir du degré d'affinité. Plus précisément, pour calculer cette fonction, on associe à chaque point représentant une clique une gaussienne dont l'amplitude est proportionnelle au degré d'affinité du nom avec la clique en question. La fonction potentielle est égale à la somme de ces gaussiennes (au signe près, puisque dans la représentation que nous avons choisie, ce sont les minima de la fonction potentielle qui représentent les sens pertinents). Autrement dit, les bassins de la fonction potentielle correspondent à des régions où l'on a une forte densité de cliques présentant beaucoup d'affinité avec le nom.

On a représenté figure 6 cette fonction potentielle pour les noms *foin*, *ton*, *homme* et *cœur*.



(5a) *foin sec* (5b) *ton sec*



(5c) *homme sec* (5d) *cœur sec*

Figure 6: *Fonctions potentielles associées à différents noms*

Différents cas de figure interprétatifs sont ainsi obtenus : pour *foin*, un sens précis, centré sur le sens 1 (‘qui ne contient pas d’eau’), pour *ton*, une indétermination entre les sens 4 et 5 (‘sévère’ et ‘brusque’), et pour *homme* une ambiguïté-alternative entre le sens 2 (‘maigre’) et 4 (‘insensible’). Le cas de *cœur* est particulièrement intéressant : La fonction potentielle a deux minima, l’un centré sur la clique {*aride, rébarbatif, sec, sévère*} et l’autre sur {*égoïste, indifférent, insensible, sec*}. Cela signifie que seuls les sens psychologiques de *sec* sont sélectionnés dans ce contexte. Dans la plupart des approches classiques, ce cas soulèverait un problème difficile, puisque le sens de *cœur* et celui de *sec* serait traités de « métaphoriques » dans cet exemple. Par exemple, dans le cadre de la théorie de Pustejovsky [PUS 95], il serait difficile d’expliquer par la coercition de type (cf. § 2.3) que les deux unités passent en même temps d’un sens premier physique à un sens dérivé psychologique. Nous ne sommes bien sûr pas confrontés à ce type de problèmes, puisque nous n’avons pas à dériver les sens dits métaphoriques à partir d’un sens dit premier. Dans notre approche, la construction d’un sens, quel qu’il soit, est toujours le résultat d’un même mécanisme dynamique.

4.2. *Constructions syntaxique et classes de sélection distributionnelle*

La méthode de calcul du sens présentée ci-dessus permet donc, en théorie, de répondre à l'objectif de désambiguïsation d'une unité polysémique. Un première évaluation quantitative, menée avec l'adjectif *sec* et vingt noms parmi les plus fréquents épithètes de *sec*, a donné des résultats remarquables : près de 80% de bonnes réponses du système quand on cherche à classer différents synonymes de *sec* comme plus ou moins acceptables en présence des noms en

question [VEN 02], [VEN 04]. Néanmoins, la généralisation de cette technique à toute sorte d'unités polysémiques dans des énoncés tout venant se heurte à un certain nombre de difficultés sur lesquelles nous sommes en train de travailler. Nous allons donc exposer chacun de ces problèmes et présenter les voies de recherche que nous explorons pour les résoudre.

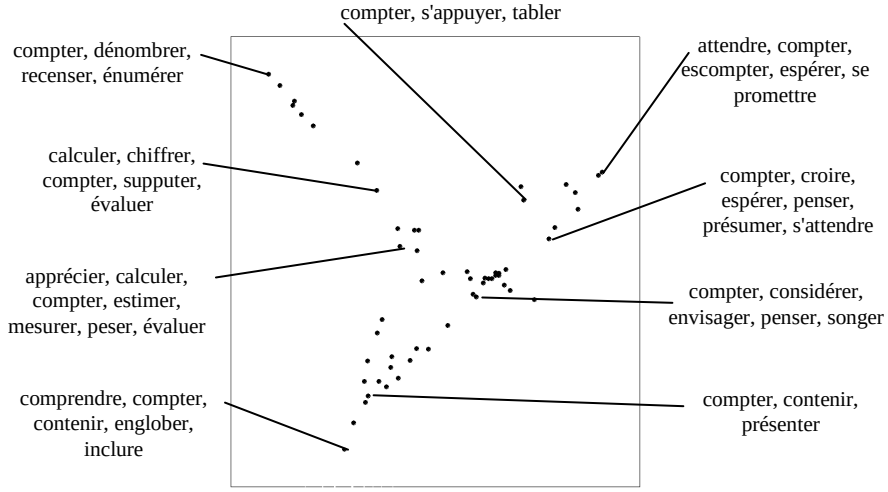
Une première série de difficultés a trait à l'analyse syntaxique de l'énoncé dans lequel se trouve l'unité polysémique étudiée. En choisissant de travailler d'abord sur des adjectifs, nous avons évité ces difficultés. En effet, le sens de l'adjectif dépend principalement du nom régissant qu'il qualifie, et la relation qui les lie est des plus simples à repérer dans un corpus : des techniques extrêmement grossières suffisent pour trouver le nom dont l'adjectif est épithète avec un taux d'erreur suffisamment faible pour ne pas trop perturber les calculs. Il n'en va pas de même pour d'autres unités polysémiques comme les verbes, par exemple. Le sens d'un verbe polysémique peut être influencé par des éléments qui entretiennent des relations syntaxiques variées avec lui, et le repérage de ces relations réclame des moyens beaucoup plus lourds d'analyse syntaxique.

Pour mener à bien l'analyse syntaxique nécessaire, nous avons utilisé SYNTEX, l'analyseur développé par Bourigault[23] [BOU 00]. Ce logiciel permet d'obtenir, entre autres, l'ensemble des relations syntaxiques d'un verbe avec ses actants (sujet, objet, compléments prépositionnels) dans un corpus tout venant avec un taux d'erreur non négligeable mais tout à fait acceptable pour notre méthode. Grâce à SYNTEX, nous avons pu vérifier que notre méthode ne s'appliquait pas uniquement à des éléments lexicaux du co-texte (comme c'était le cas avec les noms pour déterminer le sens d'un adjectif), mais aussi à des constructions syntaxiques [JAC 04], [JAC 05a]. Dans l'esprit de la théorie des grammaires constructionnelles [GOL 95], on peut en effet considérer que les constructions grammaticales sont elles-mêmes porteuses de sens, et qu'elles interagissent avec le potentiel sémantique général d'un verbe, au même titre que les éléments lexicaux cotextuels. C'est ainsi qu'un même verbe peut prendre des sens différents suivant sa construction alors que les mêmes éléments lexicaux sont utilisés : comparer par exemple *jouer un cheval* et *jouer avec un cheval*, où le verbe *jouer* change radicalement de sens (*parier* versus *s'amuser*) avec la construction. A titre d'illustration, on a présenté figure 7 les résultats obtenus avec notre méthode pour deux constructions du verbe *compter*.

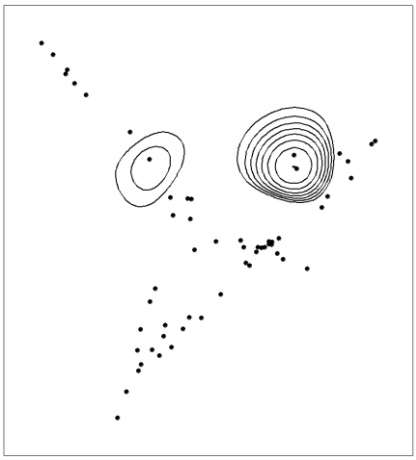
Comme on peut le constater sur la figure, les deux constructions définissent des zones différentes de l'espace sémantique de *compter*. On notera en particulier que si la construction transitive n'est pas très sélective (on trouve en effet plusieurs sens de *compter* dans cette construction : *compter les moutons, ce village compte 2000 âmes, sans compter les enfants, compter ses sous,* etc.), elle exclut tout de même une bonne partie de l'espace sémantique (notamment les sens 'importer', 'considérer', 'espérer' : *il compte beaucoup pour moi, je compte aller le voir,* etc.) ; quant à la construction *compter sur,* elle sélectionne presque exclusivement le sens 'tabler' (*compter sur quelqu'un*), avec une petite incursion du côté du dénombrement (à l'œuvre dans l'emploi *compter sur ses doigts*).

23. Tous nos remerciements à Didier Bourigault, qui a mis à notre disposition les résultats de l'analyse de SYNTEX pour les différents corpus que nous lui avons soumis.

On peut bien entendu combiner les contraintes provenant d'une construction donnée et celles d'un élément lexical, de manière à différentier par exemple *compter sur quelqu'un* et *compter sur ses doigts*. Mais on tombe alors sur une deuxième série de difficultés. D'abord, dans de nombreux cas, la méthode ne peut pas fonctionner à cause de la faible fréquence de certaines de ces combinaisons. Comme on l'a vu au § 4.1, pour calculer les degrés d'affinité à la base de la détermination de la fonction potentielle, on mesure les co-occurrences de l'élément co-textuel considéré avec tous les synonymes de l'unité polysémique étudiée. Par exemple, si l'on veut déterminer le sens de *jouer* dans *jouer du piano*, il faut chercher dans le corpus les cooccurrences de *piano* avec les synonymes de *jouer* : *pratiquer, s'amuser, parier, représenter, imiter,* etc. Pour que les calculs aient du sens, il faut évidemment que les unités soient suffisamment nombreuses dans le corpus pour que l'absence de telle ou telle combinaison soit pertinente, et que les fréquences relatives de ces cooccurrences soient significatives. Avec un corpus suffisamment gros, pour un mot comme *piano*, on peut espérer trouver suffisamment d'occurrences de *faire du piano, pratiquer le piano*, etc., pour que cela soit le cas. En revanche, avec des mots moins fréquents comme *luth* ou *ocarina*, les fréquences sont si faibles que les calculs n'ont plus aucun sens. On remarquera que l'on avait évité cette difficulté dans l'expérimentation menée sur *sec* en choisissant justement des noms très fréquents régissant cet adjectif.



(a) Espace sémantique associé à *compter*



(b) Fonction potentielle associée à la construction *compter surSN*

(c) Fonction potentielle associée à la construction *compter SN*

Figure 7: *Exemples de fonctions potentielles associées à des constructions*

La deuxième difficulté apparaît aussi clairement sur l'exemple de *jouer du piano*. Deux verbes ne se construisent pas forcément de la même manière dans des emplois où ils sont synonymes. Ainsi si l'on dit *faire du piano* (même construction que *jouer*), on dit *pratiquer le piano*, la construction transitive correspondant alors à la construction prépositionnelle de *jouer*. Ce problème, là encore, ne se posait pas pour *sec* et pour les adjectifs en général (à l'exception de la position antéposée ou postposée de l'épithète, qui peut effectivement distinguer dans certains cas des sens différents, comme dans *un vieil ami* versus *un ami vieux*, cf. [FRA 03]). Il faut donc pouvoir faire correspondre automatiquement telle construction d'un premier verbe à telle autre construction d'un deuxième verbe, ce qui implique non seulement des mesures de fréquences

significatives mais aussi de travailler sur plus d'une unité lexicale à la fois : c'est pour l'ensemble des instruments de musique que la construction *jouer de* est synonyme de *pratiquer* transitif.

Enfin, un dernier problème, plus général, vient perturber les calculs. Il s'agit de la polysémie des éléments co-textuels eux-mêmes. En effet, dans la plupart des énoncés, il n'y a pas que l'unité étudiée qui est polysémique, ce qui fait que le processus de construction du sens est plus complexe que ce que nous avons laissé entendre en nous centrant sur la détermination du sens d'une seule unité à la fois. En fait, il s'agit d'une dynamique d'interaction entre unités : chaque unité contribue à la détermination du sens des autres unités polysémiques en même temps que ces unités contribuent à déterminer le sien[24]. Considérons par exemple les expressions *jouer un morceau* et *jouer un air*. Les noms *morceau* et *air* sont eux-mêmes très polysémiques, ce qui pose au premier abord un problème de circularité : comment désambiguïser *jouer* si l'on ne connaît pas le sens exact de ces mots, sens qu'ils acquièrent justement parce qu'ils sont compléments de *jouer* ?

On peut résoudre toute cette série de difficultés à l'aide d'un seul et même procédé. Plutôt que de chercher à calculer l'influence d'une unité lexicale isolément, l'idée consiste à traiter en bloc toute la classe sémantique argumentale à laquelle elle appartient. Ainsi l'on ne cherchera pas à traiter séparément *jouer de + ocarina* ou *jouer de + luth*, mais on regroupera toutes les constructions de la forme *jouer de* + <instrument de musique>. Clairement, cela résout le problème des faibles fréquences, puisqu'en travaillant sur toutes ces unités à la fois, on a toutes les chances d'obtenir des nombres d'occurrences suffisamment élevés pour que les calculs soient significatifs. Cela permet aussi de résoudre le problème du repérage des constructions homologues de verbes synonymes, quand les deux verbes utilisent des constructions différentes pour produire le même sens, comme dans *pratiquer le piano* (construction transitive) et *jouer du piano* (construction prépositionnelle avec *de*). En effet, comme on l'a dit, ces correspondances entre constructions de verbes différents se font justement sur toute une classe d'arguments. Enfin, on peut aussi espérer limiter très sensiblement, si ce n'est éliminer totalement, les difficultés dues à la polysémie des noms en position argumentale. En effet, si l'on arrivait à ranger *morceau* et *air* dans une classe qui contient aussi *chanson*, *mélodie*, *ritournelle*, *berceuse*, etc., cela permettrait de calculer correctement le sens de *jouer*, les effets de la polysémie individuelle de tel ou tel terme étant amoindris par la prise en compte de la classe toute entière.

Reste à savoir comment construire ces classes sémantiques argumentales. De nombreux travaux ont été faits sur la construction de classes sémantiques en corpus (cf. [HAB 97], [GRE 94], [HAB 96], [AUS 00], [HAB 04]). Les classes que nous cherchons à obtenir s'apparentent d'assez près aux *classes d'objets* décrites par Gaston Gross [GRO 04]. Notamment nous partageons l'idée que « la mise au point du sens exige que l'on soit à même de préciser la nature sémantique des arguments que prend un emploi donné de prédicat ». Mais la différence, c'est que nous ne cherchons pas à obtenir des classes sémantiques générales, valables pour tout le lexique. Autrement dit, nous ne cherchons pas à construire une ontologie. Notre but est plus modeste, et la tâche, du coup, plus réaliste : il faut simplement classer, parmi les arguments d'un verbe donné dans une construction donnée, ceux qui relèvent d'un même paradigme sémantique.

24. On trouvera dans [VIC 96], chap. 8, une description plus détaillée du modèle complet, ainsi que le principe d'implémentation du modèle à l'aide de réseaux connexionnistes de taille variable. Comme on va le voir, la méthode que nous sommes en train de mettre au point devrait permettre d'éviter le recours à ces réseaux, qui se sont avérés difficiles à mettre en œuvre à grande échelle.

Ainsi nous ne cherchons pas à construire une fois pour toute la classe des instruments de musique, celle des parties du corps ou celle des qualités psychologiques. Nous cherchons simplement à savoir si dans la construction *jouer de* on doit placer *ocarina* du côté de *piano* et *violon*, ou du côté de *coude* (*jouer des coudes* se rencontre assez fréquemment), ou encore du côté de *charme* (*jouer de son charme*). En revanche, dans le contexte *porter un ocarina*, il s'agira de savoir si *ocarina* doit être classé avec *colis*, ou avec *manteau*, ou encore avec *regard* ou *coup*… Il s'agit donc de classes « en contexte », que nous avons appelées *classes de sélection distributionnelle*.

Pour obtenir automatiquement ces classes, nous avons utilisé une technique d'analyse distributionnelle sur un gros corpus[25] analysé par SYNTEX. Le principe du calcul est le suivant[26]. A chaque unité lexicale on associe sa *fiche distributionnelle*, constituée par la fréquence relative de ses différents *contextes syntactico-lexicaux* (un contexte de ce type est composé de la donnée d'un élément lexical du cotexte et de la relation qui le lie à l'unité lexicale considérée). Ce travail est effectué une fois pour toutes : chaque unité lexicale du français est donc caractérisée par sa fiche distributionnelle. Pour obtenir les classes de sélection distributionnelle associée à une construction d'un verbe donné (par exemple *jouer de*), on établit d'abord la liste des unités lexicales qui ont été rencontrées dans cette construction, puis on construit automatiquement des regroupements au sein de cette liste, en calculant une distance entre ces unités à partir de leur fiche distributionnelle[27]. Les premiers résultats obtenus [JAC 05b] par cette méthode semblent extrêmement prometteurs. Outre les résultats escomptés sur les exemples que nous avons déjà présentés comme *jouer de l'ocarina*, il faut souligner que l'on obtient d'excellents résultats sur les noms propres, qui cristallisent les différentes difficultés que nous avons évoquées. C'est ainsi que le système classe correctement les différents compléments de *descendre* dans les exemples suivants : *descendre le Gange*, *descendre Chirac*, et *descendre le Mont Blanc*. Le Gange est classé correctement dans une classe qui contient les mots *fleuve* et *rivière*, tandis que Chirac se retrouve avec les mots *homme* et *personne*, et Mont Blanc avec *piste* et *montagne*. Qui plus est, un même mot, dans différentes constructions, est classé différemment. C'est ainsi que la classe de *Wimbledon* dans *jouer à Wimbledon* contient principalement des noms de sport (elle est composée de *basket, football, loterie, tennis, rugby, jeu video, base-ball, cricket, golf,* et *loto*), alors que dans *revenir de Wimbledon*, elle est constituée de noms de lieux et d'activités (*Allemagne, Etats-Unis, guerre, mission,* et *travail*), et dans *Wimbledon décide*, elle est faite de noms d'institutions territoriales (*Europe, France, Italie, monde, pays, région,* et *ville*).). De la même manière, la classe de *morceau* dans *jouer un morceau* est composée de *chanson, pièce, instrument* alors que dans *couper en morceau*, elle est composée de *rondelle, tranche*. Ces derniers exemples montrent que l'on peut effectivement désambiguïser un nom grâce à cette méthode, comme nous l'espérions.

25. Le corpus est constitué des articles du journal *Le Monde* sur dix ans. Il a été confectionné par Benoît Habert, que nous remercions d'avoir bien voulu le mettre à notre disposition.

26. On trouvera tous les détails dans [JAC 05b].

27. Comme pour les cliques (cf. §.3.2), c'est la métrique du χ^2 qui s'est avérée donner les meilleurs résultats. Les regroupements sont obtenus par une méthode classique de *clustering* : on cherche à minimiser les distances entre les unités à l'intérieur d'un même groupe et à maximiser les distances entre groupes.

4.3. *Conclusion*

Ainsi, le modèle que nous venons de présenter montre que la polysémie lexicale, celle qui ne peut pas être traitée par des règles parce qu'elle est intrinsèque à chaque unité lexicale (contrairement aux polysémies systématiques dont nous avons parlé au § 2.3) peut effectivement être représentée et calculée en contexte, dans le cadre d'un modèle continu. Les derniers développements montrent d'ailleurs que certains aspects de la polysémie systématique peuvent aussi être capturés par le calcul des classes de sélection distributionnelle (cf. l'exemple de Wimbledon, que nous venons de présenter).

Indépendamment de la tâche de désambiguïsation lexicale, qui est notre premier objectif dans le développement de ce modèle, on peut noter que notre méthode de construction de classes sémantiques « à la volée » peut sans doute être utilisée dans d'autres tâches d'analyse sémantique d'un texte. Notamment, comme on le verra dans le chapitre suivant, la résolution des anaphores réclame, dans certains cas, de connaître la classe sémantique à laquelle peut appartenir un argument d'un verbe donné. Par exemple, dans les phrases suivantes : *La voiture a renversé Marie. Elle a été blessée assez gravement.*, il faut absolument savoir que le sujet de *être blessé* doit être un humain ou un animal pour trouver que l'antécédent de *Elle* est *Marie* et non pas *la voiture.* Notre méthode devrait pouvoir être appliquée avec profit à ce genre de problèmes en fournissant les classes de sélection distributionnelles pertinentes chaque fois que l'on en a besoin.

En ce qui concerne la désambiguïsation lexicale, quand ce modèle sera complètement opérationnel (rappelons encore une fois que ce travail n'est pas encore terminé), on peut penser qu'il sera d'une grande utilité pour de nombreuses applications en traitement automatique des langues. Il faut souligner que ce que nous avons appelé « désambiguïsation » ici, c'est la capacité pour le système d'exhiber des synonymes pertinents de l'unité étudiée dans le contexte étudié. C'est exactement ce dont on a besoin dans certaines applications en recherche d'information, par exemple. Dans d'autres cas, on peut avoir besoin d'un outil de désambiguïsation qui fournisse autre chose qu'une liste de synonymes pertinents de l'unité dont on a calculé le sens. Par exemple, dans des tâches d'Extraction d'Information, ce dont on a besoin, c'est de savoir si l'unité lexicale que l'on a repérée a bien le sens qui nous intéresse, c'est-à-dire le sens attendu dans le formulaire que l'on doit remplir (cf. chap. 8). Il reste donc dans ce cas une partie importante du travail à effectuer pour utiliser ce modèle : il faut caractériser le sens recherché, autrement dit spécifier la région de l'espace sémantique associé à l'unité qui correspond à ce sens. L'avantage du modèle, c'est qu'il permet différents degrés de finesse suivant les applications. Si l'on n'a besoin que d'une désambiguïsation grossière, on peut séparer l'espace sémantique de l'unité considérée en quelques grandes zones auxquelles correspondent des ensembles de sens très différents. En revanche, si l'on a besoin de plus de précision, on peut mettre en place une correspondance plus fine, jusqu'au niveau le plus détaillé du modèle, à savoir la clique individuelle. Même si cela reste pour l'instant hypothétique par bien des aspects, cela montre que le problème de la polysémie lexicale, à condition qu'on lui accorde l'attention qu'il mérite, doit pouvoir être traité de manière satisfaisante, quel que soit le niveau d'exigence que l'application exige.

5. Bibliographie

[AUS 00] AUSSENAC-GILLES N., BIÉBOW B., SZULMAN N., « Revisiting Ontology Design: a method based on corpus analysis », *Actes de 12th International Conference on Knowledge Engineering and Knowledge Management*, Juan-Les-Pins, 2000.

[BER 70] BERGE C., *Graphes et hypergraphes*, Paris, Dunod, 1970.

[BOU 00] BOURIGAULT D., FABRE C., « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de Grammaires,* n° 25, pp. 131-151, 2000.

[BOU 01] BOURIGAULT D., JACQUEMIN Ch., L'HOMME M.-C., *Recent Advances in Computational Terminology*, Amsterdam, John Benjamins, 2001.

[CAD 01] CADIOT P., VISETTI Y.-M., *Pour une théorie des formes sémantiques, Motifs, profils, thèmes*, Paris, PUF, 2001.

[CRO 04] CROFT W., CRUSE D.A., *Cognitive Linguistics*, Cambridge University Press. 2004.

[CRU 00] CRUSE D.A., « Aspects of the microstructure of word meanings », in RAVIN Y. , LEACOCK C. (eds), *Polysemy: theoretical and computationl approaches*, Oxford University Press, p. 30-51, 2000.

[CUL 90] CULIOLI A., *Pour une linguistique de l'énonciation*, Paris, Ophrys, 1990.

[DES 85] DESCLES, J.-P., « Représentation des connaissances : archétypes cognitifs, schèmes conceptuels, schèmes grammaticaux », *Actes Sémiotiques*, Documents(VII), p. 69-70, 1985.

[DUV 02] DUVIGNAU K., La métaphore, berceau et enfant de la langue. La métaphore verbale comme approximation sémantique par analogie dans les textes scientifiques et les productions enfantines (2-4 ans), Thèse en Sciences du Langage, Université Toulouse Le-Mirail, 2002.

[DUV 03] DUVIGNAU, K., « Métaphore verbale et approximation », in DUVIGNAU, K., GASQUET, O., GAUME, O. (eds) *Regards croisés sur l'analogie*. *Revue d'Intelligence Artificielle*, spécial, Vol 5/6. Hermès Sciences, p. 869-881, 2003.

[FAU 84] FAUCONNIER G., *Les espaces mentaux*, Editions de Minuit, 1984.

[FEL 98] FELLBAUM C., *Wordnet: an Electronic Lexical Database*, Cambridge, MIT Press, 1998.

[FRA 03] FRANÇOIS J., MANGUIN J.L. VICTORRI B., « La réduction de la polysémie adjectivale en cotexte nominal: une méthode de sémantique calculatoire », *Cahiers du Crisco,* 14, http://www.crisco.unicaen.fr, 2003.

[FRA 05] FRANÇOIS J., VICTORRI B., MANGUIN J.L., « Polysémie adjectivale et synonymie: l'éventail des sens de curieux », in SOUTET O. (ed.) *La polysémie*, Presses de l'Université de la Sorbonne, 2005.

[GAU 02] GAUME B., DUVIGNAU K., GASQUET O., GINESTE M-D., « Forms of Meaning, Meaning of Forms », *Journal of Experimental and Theoretical Artificial Intelligence*, 14(1), p. 61-74, 2002.

[GAU 03] GAUME B., « Analogie et Proxémie dans les réseaux petits mondes », in DUVIGNAU, K., GASQUET, O., GAUME, O. (eds) *Regards croisés sur l'analogie*. *Revue d'Intelligence Artificielle*, Vol 5/6. Hermès Sciences, 2003.

[GOL 95] GOLDBERG A., *Constructions: a construction grammar approach to argument structure*, Chicago and London, University of Chicago Press, 1995.

[GRE 94] GREFENSTETTE G., *Explorations in Automatic Thesaurus Discovery*, London, Kluwer Academic Publishers, 1994.

[GRO 04] GROSS G., « Réflexions sur le traitement automatique des langues », *Actes de JADT 2004*, Vol. 1, 545-556, 2004.

[HAB 96] HABERT B., NAZARENKO A., « La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience », *Journées sur l'acquisition des connaissances*, AFIA, Sète, 1996.

[HAB 97] HABERT B., NAZARENKO A., SALEM, A., *Les linguistiques de corpus*, Armand Colin, 1997.

[HAB 04] HABERT B., ILLIOUZ G., FOLCH H., « Dégrouper les sens: pourquoi, comment? », *Actes de JADT 2004*, Vol. 1, 565-576, 2004.

[IDE 90] IDE N., VÉRONIS J., « Word Sense Disambiguation with very large neural networks extracted from machine readable dictionaries », *Proceedings of the 14th International Conference on Computational Linguistics*, 1990.

[JAC 04] JACQUET G., « Using the construction grammar model to disambiguate polysemic verbs in French », *Actes de ICCG3 (Third International Conference on Construction Grammar)*, 2004.

[JAC 05a] JACQUET G., « A model of disambiguation of polysemic verbs in French », *Constructions*, http://www.constructions-online.de/, 2005.

[JAC 05b] JACQUET G. et VENANT F., « Construction automatique de classes de sélection distributionnelle », *Actes du colloque TALN*, 2005.

[KLE 90] KLEIBER G., *La sémantique du prototype : Catégories et sens lexi*cal, Paris, P.U.F., 1990.

[KLE 94] KLEIBER G., *Nominales : essai de sémantique référentielle*, Paris, Armand Colin, 1994.

[KLE 99] KLEIBER G. *Problèmes de sémantique : la polysémie en questions*, Paris, P.U.F., 1999.

[LAK 87] LAKOFF G., *Women, Fire and Dangerous Things*, University of Chicago Press, 1987.

[LAN 99] LANGACKER R. W., *Foundations of Cognitive Grammar*, vol. 1 : *Theoretical Prerequisites*, Stanford University Press, 1987.

[MAN 99] MANGUIN J.L., VICTORRI B., « Représentation géométrique d'un paradigme lexical », *Actes de la 8ème conférence TALN*, vol. 1:363-368, 1999.

[MAN 01] MANGUIN J.L., « Construction d'espaces sémantiques associés aux verbes de déplacement d'objets à partir des données des dictionnaires informatisés des synonymes », *Syntaxe et Sémantique*, 2:287-300, 2001.

[MAN 02] MANGUIN J.L., FRANÇOIS J., VICTORRI B., « Polysémie adjectivale et rection nominale: quand gros et gras sont synonymes », in FRANÇOIS J. (ed.), *L'adjectif en français et à travers les langues*, Presses Universitaires de Caen, 2002.

[NUN 78] NUNBERG G., *The pragmatics of reference*, Indiana University Linguistics Club, 1978.

[NUN 95] NUNBERG G. « Transfers of Meaning », *Journal of Semantics*, 17, 109-132, 1995.

[NUN 97] NUNBERG G., ZAENEN A., « La polysémie systématique dans la description lexicale », *Langue Française*, 113, 12-23, 1997.

[PIC 86] PICOCHE J., *Structures sémantiques du lexique français*, Nathan, 1986.

[PIE 00] PIERREL J.M. (ed.), *Ingénierie des langues*, Paris, Hermès, 2000.

[PIT 85] PITRAT J., *Textes, ordinateurs et compréhension*, Eyrolles, 1985.

[PLO 98] PLOUX S., VICTORRI B., « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *Traitement automatique des langues*, 39/1, p.161-182, 1998.

[PLO 03] PLOUX S. & JI H., « A model for matching semantic maps between languages (French/English, English/French) », *Computational Linguistics* 29(2), p. 155-178, 2003.

[PUS 95] PUSTEJOVSKY J., *The generative lexicon*, Cambridge, MIT Press, 1995.

[RAS 94] RASTIER F., CAVAZZA M., ABEILLE A., *Sémantique pour l'analyse – de la linguistique à l'informatique*, Masson, 1994.

[REI 77] REINGOLD E.M., NIEVERGELT J., DEO N., *Combinatorial Algorithms, Theory and Practice*, Prentice-Hall, 1977.

[RIC 75] RICOEUR P., *La métaphore vive*, Seuil, 1975.

[SAD 86] SADOCK J.M., « Vagueness as a vague concept », *Quaderni di Semantica*, 7-2, 1986.

[VEN 02] VENANT F., Polysémie adjectivale et calcul du sens, mémoire de DEA de Sciences Cognitives, Paris, EHESS, 2002.

[VEN 04] VENANT F., « Polysémie et calcul du sens », in *Le poids des mots*, *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles* (*JADT*), 2004.

[VIC 96] VICTORRI B., FUCHS C., *La polysémie – Construction dynamique du sens*, Paris, Hermès, 1996.

[VIC 97] VICTORRI B., « La polysémie : un artefact de la linguistique ? », *Revue de Sémantique et de Pragmatique*, 2, p. 41-62, 1997.

[VIC 02] VICTORRI B., « Espaces sémantiques et représentation du sens », in *Textualités et nouvelles technologies*, éc/artS, 3, 2003. [[XX éc/artS ??]]

[WAR 85] WARNESSON I., « Applied Linguistics : Optimization of Semantic Relations by Data Aggregation Techniques », Journal of Applied Stochastic Models and Data Analysis, vol.1/2:121-143., 1985.

[WIT 53] WITTGENSTEIN L., *Philosophical Investigations*, New York, Macmillan, 1953.